POLI 7050

Spring 2008

February 27, 2008

Unordered Response Models I

Introduction

For the next couple weeks we'll be talking about unordered, polychotomous dependent variables. Examples include:

- Voter choice (e.g. Bush/Clinton/Perot, Labour/liberal/Conservative, etc.)
- Voting in legislatures (e.g. yea/nay/prestent).
- Political strategy by interest groups (choice to lobby/litigate/protest).
- Decisions by nations to join one of several international organizations.
- Economics: Occupational choice, marketing (choosing brands of tuna), etc.
- Others?

Q: Where do variables like this come from?

- Some are "natural" types (e.g. discrete choices)...
  - Votes for Presidential candidates.
  - Elazar's "types" of states (individualistic, moralistic, traditionalistic).
- Others are "aggregates":
  - For example: Famous study by Schmidt and Strauss (1975) on occupational attainment:
    - They wanted to explain occupational choice, BUT there are lots of job types.
    - Divided up types of jobs: menial, blue-collar, craft, white-collar, and professional.
  - Or, we might classify international interactions into economic, political, military, etc.
  - Obviously, we need to be sensitive to classification schemes when making such classifications.
    - Different classifications = different results.
    - May consider some sort of data reduction procedure first (e.g. factor analysis).
    - Techniques also offer some tests for whether classification schemes are OK or not.

ALSO: What about ordered categories?

- If the categories are ordered, using an unordered model has its ups and downs.
  - Good: Allows you to get around the "parallel regression" idea.
  - In other words, the independent variables can have different effects across different categories.
  - Bad: Is *inefficient* (doesn't use information about the ordering of the categories).
- Conversely, if you use ordered probit/logit on unordered categories, you get garbage.
- One option is to do it both ways, and compare estimates, etc. (more on this later).

Multinomial choice models are being much more widely used...

- Folks are starting to realize that simply dichotomizing polychotomous variables can be very misleading.
- E.g., the Whitten and Palmer article:
  - Voter choice in multi-party elections needs to be modeled as what it is: a choice across a range of options.
  - These models let you extract additional information out of your data.
- The software is also readily available (more on this in a bit as well).

Today we'll talk about the *multinomial logit* ("MNL") and "conditional logit" ("CL") models; though, as we'll see, they're really the same model.

Motivation(s)

There are several ways of motivating the multinomial logit model...

As A Model of Discrete Outcomes

This is the motivation Long takes in his book. Consider once again a dataset of N individuals $i = \{1, 2, ...N\}$, and a dependent variable Y_i which can take on any of J *unordered* values.

- Call $\Pr(Y_i = j) = P_{ij}$, and note that, by definition, $\sum_{j=1}^{J} P_{ij} = 1$ (that is, every observation has one realization of the J possible outcomes).
- We want to allow the probability of $Y_i = j \in J$ to vary as a function of some k independent variable(s) $\mathbf{X}_i$, indexed by a $k \times 1$ vector of parameters specific to that outcome $\boldsymbol{\beta}_j$.

- To insure that each of these probabilities are positive, we use the exponential function:

$$P_{ij} = \exp(\mathbf{X}_i \boldsymbol{\beta}_j)$$

- But now $\sum_{j=1}^{J} P_{ij} \neq 1$, which (by definition) it has to. So...
- ...we *rescale* the P_{ij}s, by dividing each by the sum of all the P_{ij}s:

$$\Pr(Y_i = j) \equiv P_{ij} = \frac{\exp(\mathbf{X}_i \boldsymbol{\beta}_j)}{\sum_{j=1}^{J} \exp(\mathbf{X}_i \boldsymbol{\beta}_j)} \tag{1}$$

  That way, observation i's probability associated with category j is expressed as a fraction of the sum of *all* of observation i's probabilities across the various categories J. That ensures that $\Pr(Y_i = j) \in (0, 1)$ and that $\sum_{j=1}^{J} \Pr(Y_i = j) = 1.0$.

Equation (1) provides us with our primary statement of probability for the MNL model. But, this model is also *unidentified*; if we know the $(J - 1) \times k$ values of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ...\boldsymbol{\beta}_{J-1}$, then we also know the probability of choosing the remaining alternative. Put differently, this model can have an infinite number of parameters which will generate the identical set of probabilities – not unlike the situation with the constant term in the ordered probit/logit models.

We deal with this in the usual way: by imposing constraints on the parameters.

- The most common is to set the parameters for one of the J alternatives to zero (usually the "first" one, i.e., $\boldsymbol{\beta}_1$).
- This is referred to as the *baseline* (or *omitted*) category, and it provides the "reference point" for the other probabilities.
- It is the alternative to which the others are compared when the model's coefficients are estimated.
- Note that when we do this, the formula for $\Pr(Y_i = 1)$ reduces to:

$$\frac{1}{1 + \sum_{j=2}^{J} \exp(\mathbf{X}_i \boldsymbol{\beta}'_j)}$$

  and that for the other $J - 1$ alternatives are:

$$\frac{\exp(\mathbf{X}_i \boldsymbol{\beta}'_j)}{1 + \sum_{j=2}^{J} \exp(\mathbf{X}_i \boldsymbol{\beta}'_j)}$$

  where $\boldsymbol{\beta}'_j = \boldsymbol{\beta}_j - \boldsymbol{\beta}_1$ are the "rescaled" parameters (that is, they express the influence of the various $\mathbf{X}$s on $\Pr(Y_i = j)$ relative to $\Pr(Y_i = 1)$).

An Alternative Motivation: The Discrete Choice Model

Like we did with the binary choice model, consider an individual i choosing among J alternatives.

- S/he has a utility U_{ij} associated with each choice j.
- That utility has a stochastic (random) part and a systematic part; that is, $U_{ij} = \mu_i + \epsilon_{ij}$.
- The systematic part is a function of some variables associated with the *individual*, who may give different weights to those characteristics across alternatives (that is, $\mu_i = \mathbf{X}_i \boldsymbol{\beta}_j$).
- The individual chooses among the alternatives in such a way that maximizes his or her utility, so that

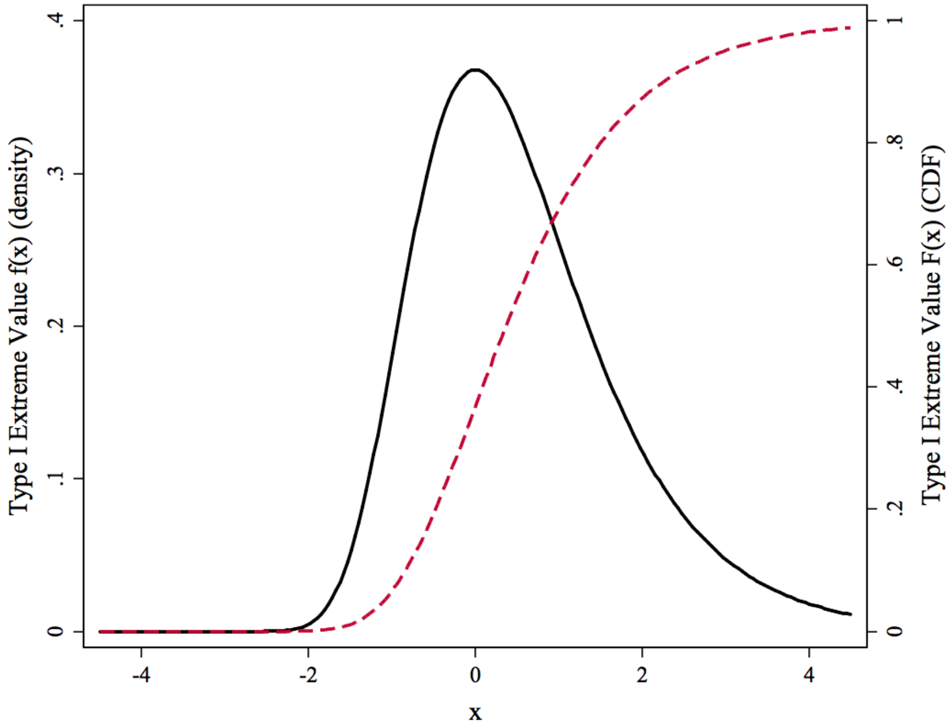
$$\begin{aligned}
\Pr(Y_i = j) &= \Pr(U_{ij} > U_{i\ell} \, \forall \, \ell \neq j \in J) \\
&= \Pr(\mu_i + \epsilon_{ij} > \mu_i + \epsilon_{i\ell} \, \forall \, \ell \neq j \in J) \\
&= \Pr(\mathbf{X}_i \boldsymbol{\beta}_j + \epsilon_{ij} > \mathbf{X}_i \boldsymbol{\beta}_\ell + \epsilon_{i\ell} \, \forall \, \ell \neq j \in J) \\
&= \Pr(\epsilon_{ij} + \mathbf{X}_i \boldsymbol{\beta}_j - \mathbf{X}_i \boldsymbol{\beta}_\ell > \epsilon_{i\ell} \, \forall \, \ell \neq j \in J) \qquad (2)
\end{aligned}$$

- In other words, *the chosen alternative is the one in which the difference between its stochastic component and that of any alternative is greater than the difference in the systematic parts.*
- This can happen either because ϵ_{ij} is large, or because $\mathbf{X}_i \boldsymbol{\beta}_j$ is large, or both...

Now, *what is the probability of (2) happening*? To answer that, we need (as we've always done so far) to specify the distribution of the error terms ϵ_{ij}.

- We *could* use a multivariate normal distribution; this would yield a *multivariate probit* model. We'll come back to this alternative a bit later.
- Instead, we choose a standard (mean-zero, variance-one) *Type I Extreme Value* Distribution (a.k.a. a *Gumbel* distribution).
  - Its density function (PDF) is $f(\epsilon) = \exp[-\epsilon - \exp(-\epsilon)]$.
  - Its distribution function (CDF) is $\int f(\epsilon) \equiv F(\epsilon) = \exp[-\exp(-\epsilon)]$.
  - The density and CDF are illustrated in Figure 1...

Figure 1: A Type I Extreme Value PDF and CDF



Now, if the errors in the random utility model are independently and identically distributed according to a Type I Extreme Value distribution, then:

$$
\begin{aligned}
\Pr(Y_i = j) &= \Pr(U_j > U_1, U_j > U_2, ...U_j > U_J) \\
&= \int f(\epsilon_j) \left[\int_{-\infty}^{\epsilon_{ij} + \mathbf{X}_i \boldsymbol{\beta}_j - \mathbf{X}_i \boldsymbol{\beta}_1} f(\epsilon_1) d\epsilon_1 \times \int_{-\infty}^{\epsilon_{ij} + \mathbf{X}_i \boldsymbol{\beta}_j - \mathbf{X}_i \boldsymbol{\beta}_2} f(\epsilon_2) d\epsilon_2 \times ... \right] d\epsilon_j \\
&= \int f(\epsilon_j) \times \exp[-\exp(\epsilon_{ij} + \mathbf{X}_i \boldsymbol{\beta}_j - \mathbf{X}_i \boldsymbol{\beta}_1)] \times \exp[-\exp(\epsilon_{ij} + \mathbf{X}_i \boldsymbol{\beta}_j - \mathbf{X}_i \boldsymbol{\beta}_2)] \times ... d\epsilon_j \\
&= \frac{\exp(\mathbf{X}_i \boldsymbol{\beta}_j)}{\sum_{j=1}^{J} \exp(\mathbf{X}_i \boldsymbol{\beta}_j)}
\end{aligned}
$$

In other words, we wind up with an MNL model.[1]

[1]Maddala (1983, pp. 60-61) offers a somewhat different derivation of this model.

Estimation

Either way we motivate the model, we can estimate it using regular maximum likelihood methods. To do this, we once again consider the joint probability of observing each of the actual Y_is given the $\mathbf{X}$s.

- The probability that $Y_i = j$ is specified in (1), above...
- As we did in the ordered-variable case, define a variable $\delta_{ij} = 1$ if $Y_i = j$ and 0 otherwise.

We can then write the likelihood for a single observation as:

$$L_i = \prod_{j=1}^{J} [\Pr(Y_i = j)]^{\delta_{ij}}. \tag{3}$$

We can take the product across all N observations to get the likelihood for all the data:

$$L = \prod_{i=1}^{N} \prod_{j=1}^{J} [\Pr(Y_i = j)]^{\delta_{ij}} \tag{4}$$

and we then take the log of this, noting the formula for $\Pr(Y_i = j)$ in (1), above:

$$\ln L = \sum_{i=1}^{N} \sum_{j=1}^{J} \delta_{ij} \ln \left(\frac{\exp(\mathbf{X}_i \boldsymbol{\beta}_j)}{\sum_{j=1}^{J} \exp(\mathbf{X}_i \boldsymbol{\beta}_j)} \right) \tag{5}$$

Notes on Estimation

Maddala (1983) shows that this likelihood is very well-behaved...

- The first and second derivatives are quite easy.
- The second derivative is everywhere negative definite, meaning that the function is globally concave.
- That means that Newton-Raphson will always converge to a global maximum (and so N-R is almost always used).

Maximization of (5) with respect to the parameters yields MLEs for the $\hat{\boldsymbol{\beta}}_j$s. Note that there are separate sets of $\hat{\boldsymbol{\beta}}$s for each of the alternatives, minus the omitted baseline category (which are implicitly restricted to equal zero). Each coefficient reflects the change in the probability of that outcome, relative to that of the omitted category, associated with a change in the independent variable. Moreover, the coefficients are related in a very specific way (as we'll see in a minute...).

An Example: Political Party and Voting in the 1992 U.S. Presidential Election

To illustrate the multinomial logit model, we'll look at some data on the 1992 Presidential election, from the ANES...

- Y is presidential vote chice, with Bush=1, Clinton=2, and Perot=3 (valid $N = 1473$).
- The lone covariate is political party identification, the usual seven-point ANES measure ranging from one (for "Strong Democrats") to seven (for "Strong Republicans").
- In the first model, the omitted category is Bush:

```
. mlogit presvote partyid, baseoutcome(1)

Multinomial logistic regression                   Number of obs   =       1473
                                                  LR chi2(2)      =     891.93
                                                  Prob > chi2     =     0.0000
Log likelihood = -1083.4749                       Pseudo R2       =     0.2916
------------------------------------------------------------------------------
    presvote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
2            |
     partyid |  -1.163168   .0545618   -21.32   0.000    -1.270107   -1.056229
       _cons |   4.842495   .2373171    20.41   0.000     4.377362    5.307628
-------------+----------------------------------------------------------------
3            |
     partyid |  -.4826514   .0475563   -10.15   0.000      -.57586   -.3894427
       _cons |   1.815236    .245572     7.39   0.000     1.333923    2.296548
------------------------------------------------------------------------------
(presvote==1 is the base outcome)
```

We'll talk more about how to interpret these results next week. For now, consider two things:

- The positive coefficient estimates indicate that, as voters get more Republican, they are (much) less likely to vote for Clinton (vs. Bush), and (somewhat) less likely to vote for Perot (again, as compared to Bush).
- Changing the "baseline" category changes the coefficients, but yields the same results. For example, if we use the `-baseoutcome()-` option to specify Clinton as the omitted (baseline) category, we get:

```
. mlogit presvote partyid, baseoutcome(2)

Multinomial logistic regression                   Number of obs   =       1473
                                                  LR chi2(2)      =     891.93
                                                  Prob > chi2     =     0.0000
Log likelihood = -1083.4749                       Pseudo R2       =     0.2916
------------------------------------------------------------------------------
    presvote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
     partyid |   1.163168   .0545578    21.32   0.000     1.056236    1.270099
       _cons |  -4.842495   .2372982   -20.41   0.000    -5.307591   -4.377399
-------------+----------------------------------------------------------------
3            |
     partyid |   .6805163   .0477682    14.25   0.000     .5868923    .7741403
       _cons |  -3.027259   .1782498   -16.98   0.000    -3.376622   -2.677896
------------------------------------------------------------------------------
(presvote==2 is the base outcome)
```

Note the differences in the coefficient estimates as compared to the first set. Because we're now comparing Bush with Clinton, the first coefficient is the same, but the sign is reversed: now increasing Republicanism results in an *increase* in the odds of voting for Bush (over Clinton). Likewise, note that the coefficient for the Clinton-Perot comparison is simply the difference between the Bush-Clinton and Bush-Perot comparisons in the first model (i.e., -0.47303 - (-1.13178) = 0.65875).

We can do a similar thing if we omit Perot as the "baseline" category:

```
. mlogit presvote partyid, baseoutcome(3)

Multinomial logistic regression                   Number of obs   =       1473
                                                  LR chi2(2)      =     891.93
                                                  Prob > chi2     =     0.0000
Log likelihood = -1083.4749                       Pseudo R2       =     0.2916
------------------------------------------------------------------------------
    presvote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
     partyid |   .4826514   .0475563    10.15   0.000     .3894427      .57586
       _cons |  -1.815236    .245572    -7.39   0.000    -2.296548   -1.333923
-------------+----------------------------------------------------------------
2            |
     partyid |  -.6805163     .04777   -14.25   0.000    -.7741438   -.5868889
       _cons |   3.027259   .1782539    16.98   0.000     2.677888     3.37663
------------------------------------------------------------------------------
(presvote==3 is the base outcome)
```

Here,

- The estimate for the Perot-Clinton comparison is just the negative of that for the previous Clinton-Perot one, and
- That for the Perot-Bush comparison is the negative of that for the Bush-Perot comparison in the first table.

Note as well that in all three models, all other statistics – the log-likelihood, the χ^2, goodness-of-fit, etc. – are identical. This is because the models and data are the same; the only thing that is changing is the reference / "baseline" category.

Multinomial Logit vs. Binary Logit

One question you might have at this point is how multinomial logit compares to binary (that is, choice-by-choice) logit. One can, for example, imagine estimating a model where each choice (=1) is contrasted with all others (=0); for example, a model of whether or not each survey respondent voted for Bush, Clinton, or Perot.

As it happens, such a series of models are effectively equivalent to the MNL we've talked about so far. We'll go into this at greater length in a few weeks, when we discuss the "independence of irrelevant alternatives" assumption (and models that relax it). For now, just recognize that the $J-1$ sets of estimates MNL yields are actually quite mathematically similar – and will thus give similar substantive inferences – to estimating $J-1$ separate binary logits.

Conditional Logit

You should understand one key thing about the conditional logit model:

It is exactly the same as the multinomial logit model. Period.

This means that:

- The formulas are very similar-looking, and can be shown to be mathematically equivalent (e.g. Long, Maddala),
- Estimation is done exactly the same way,
- You will get identical results from MNL and CL if you run the same model,
- All of the ways in which you can interpret a MNL model, you can also interpret a conditional logit.

That said, it turns out that the CL is actually more flexible than the MNL model in most circumstances. So, why the differences?

Conditional Logit: A Primer

The MNL model measures the impact of observation-specific variables on the probability of choosing one of J discrete outcomes. In the MNL model, the data look like this:

```
. list caseid presvote partyid

     +-----------------------------+
     | caseid   presvote   partyid |
     |-----------------------------|
  1. |   3001          1         6 |
  2. |   3002          1         7 |
  3. |   3003          1         7 |
  4. |   3005          2         6 |
  5. |   3006          2         2 |
     ...
```

etc. This is familiar: there is one line of data per observation, and covariates (that is, $\mathbf{X}$s) vary across individuals/cases, but not across choices. Estimation yields separate effects for each variable on each choice.

But what if we have some variable (say, $\mathbf{Z}_{ij}$) that varies across *choices*, or across *both* observations and choices? For example, in our study of the 1992 Presidential election, we might want to incorporate some measure of the voter's evaluation of each candidate, which will vary across choices/candidates. Here, these are the "feeling thermometer" scores we talked about before (designated `BushFT`, `ClintonFT`, and `PerotFT`):

```
. list caseid presvote partyid BushFT ClintonFT PerotFT

     +-------------------------------------------------------------+
     | caseid   presvote   partyid   BushFT   Clinto~T   PerotFT |
     |-------------------------------------------------------------|
  1. |   3001          1         6       85         30         0 |
  2. |   3002          1         7      100          0         0 |
  3. |   3003          1         7       85         30        60 |
  4. |   3005          2         6       40         60        60 |
  5. |   3006          2         2       30         70        50 |
     ...
```

The *conditional logit* ("CL") model allows us to estimate the effect of such choice-specific variables on the probability of choosing a particular alternative. To do this, we need to restructure the data into what we Stata calls "long" format...

```
. expand 3
(2946 observations created)

. sort caseid

. quietly by caseid : gen candid=_n

. gen vote=0

. replace vote=1 if candid==presvote
(1473 real changes made)

. gen FT=BushFT if candid==1
(2946 missing values generated)

. replace FT=ClintonFT if candid==2
(1473 real changes made)

. replace FT=PerotFT if candid==3
(1473 real changes made)

. list caseid presvote candid vote FT

     +-----------------------------------------+
     | caseid   presvote   candid   vote    FT |
     |-----------------------------------------|
  1. |   3001          1        1      1    85 |
```

```
  2. |    3001          1        2       0     30 |
  3. |    3001          1        3       0      0 |
  4. |    3002          1        1       1    100 |
  5. |    3002          1        2       0      0 |
     |-----------------------------------------|
  6. |    3002          1        3       0      0 |
      ...
```

etc. Note that:

- `caseid` still indexes the observation (i.e., survey respondent).
- `candid` indexes the (possible) choice: 1 (Bush), 2 (Clinton) or 3 (Perot).
- `presvote` is the same as before, but
- `vote` is a dummy variable, equaling 1 if that observation (respondent) chose (voted for) that alternative (candidate), and 0 otherwise.
- `FT` is that respondent's "feeling thermometer" score for that candidate; note that, unlike (say) the `partyid` variable, `FT` *changes* across alternatives.

The conditional logit estimates a model with the basic probability function:

$$\Pr(Y_{ij} = j) = \frac{\exp(\mathbf{Z}_{ij}\gamma)}{\sum_{j=1}^{J}\exp(\mathbf{Z}_{ij}\gamma)} \tag{6}$$

where $\mathbf{Z}_{ij}$ is a vector of independent variables that varies across both observations i and choice alternatives j, and γ is a $k \times 1$ vector of coefficients to be estimated.

Note a couple things:

- Because the $\mathbf{Z}$s vary across both observations and alternatives, we assume that the effect of a change in $\mathbf{Z}$ on the probability of choosing a particular alternative is constant across alternatives.
- Hence, we only estimate k parameters (not $(J-1)k$, as in the MNL model).

So, the conditional logit model is useful when we want to estimate the impact of alternative- or choice-specific variables on a polychotomous dependent variable. E.g.:

- Choices of transportation alternatives, according to how long each alternative takes, or
- Choosing political candidates on the basis of where each respondent places that candidate ideologically, relative to him or her; or "feeling thermometers"...

Unsurprisingly, the estimation results look a fair bit like that for a binary logit:

```
. clogit vote FT, group(caseid)

Conditional (fixed-effects) logistic regression   Number of obs   =       4419
                                                  LR chi2(1)      =    1600.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -817.96837                       Pseudo R2       =     0.4945
------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          FT |   .0766544   .0031035    24.70   0.000     .0705716    .0827372
------------------------------------------------------------------------------

. logit vote FT

Logistic regression                               Number of obs   =       4419
                                                  LR chi2(1)      =    1777.33
                                                  Prob > chi2     =     0.0000
Log likelihood = -1924.0926                       Pseudo R2       =     0.3159
------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          FT |   .0744263   .0024393    30.51   0.000     .0696453    .0792073
       _cons |   -5.00402   .1566234   -31.95   0.000    -5.310997   -4.697044
------------------------------------------------------------------------------
```

Modeling Observation- and Choice-Specific Effects

Question: What happens if we want to estimate joint effects of variables which are constant across choices (e.g. demographic effects) and those which vary across alternatives (e.g. candidate distance)?

- Suppose we think, say, that both party identification and "feelings" affect candidate choice, or
- that a Congressperson's decision to run for reelection, or seek some higher office, depends on both individual factors (e.g. his/her age) and the expected likelihood of winning in each case (which varies by the office he's aspiring to, as well as by candidate).

The simplest kind of alternative-specific effect is simply a *fixed effect*...

- The CL model in the form given above doesn't include a constant term...
- We might (probably would) expect some alternatives to be more or less likely than others, just in general...

- So, we can include indicator variables for each possible alternative (a la “fixed effects” for each alternative).

In fact, we can then use these fixed effects to “trick” the CL into estimating a model with independent variables which do not vary across alternatives. To understand why this is, think about the notion of an interaction term with a dummy variable...

- Term allows the effect of some variable to have a different effect on the dependent variable for each of two groups.
- E.g. ideology might have a different effect on judicial voting for activist judges than it does for restraintists.
- Here, we can use this idea:
  - The interaction of an observation-specific variable with an alternative-specific constant allows the observation-specific variable's effect to be different for that alternative than for the others.
  - This is how we can get the CL model to estimate effects for observation-specific variables

It turns out that, if we do this, its possible (in fact, easy) to get results identical to those you'd get from running a MNL model...

MNL and CL: An Example

Returning to our 1992 election data, we'll start by replicating the original multinomial logit model. Note that, because we now have three lines of data for each observation, we estimate the model only for those lines of data for which `candid==1` (we could have used `candid==2` or `3` and the results would be the same):

```
. mlogit presvote partyid if candid==1, baseoutcome(1)

Multinomial logistic regression                   Number of obs   =       1473
                                                  LR chi2(2)      =     891.93
                                                  Prob > chi2     =     0.0000
Log likelihood = -1083.4749                       Pseudo R2       =     0.2916
------------------------------------------------------------------------------
    presvote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
2            |
     partyid |  -1.163168   .0545618   -21.32   0.000    -1.270107   -1.056229
       _cons |   4.842495   .2373171    20.41   0.000     4.377362    5.307628
-------------+----------------------------------------------------------------
```

```
3            |
     partyid |  -.4826514   .0475563   -10.15   0.000      -.57586   -.3894427
       _cons |   1.815236    .245572     7.39   0.000     1.333923    2.296548
------------------------------------------------------------------------------
(presvote==1 is the base outcome)
```

As we noted before, `partyid` has a negative, significant effects on the probabilities of choosing Clinton or Perot (versus Bush). Similarly, if we wanted to know the effect of `FT` on choosing each of the various alternatives, we'd use `-clogit-`:

```
. clogit vote FT, group(caseid)

Conditional (fixed-effects) logistic regression   Number of obs   =       4419
                                                  LR chi2(1)      =    1600.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -817.96837                       Pseudo R2       =     0.4945
------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          FT |   .0766544   .0031035    24.70   0.000     .0705716    .0827372
------------------------------------------------------------------------------
```

Here, `FT` has a (unsurprisingly) significant, positive impact on the probability of choosing that candidate; as `FT` increases for a particular choice, the probability of its being chosen also increases.

Of course, what we're really interested in is often the effect of, say, `FT` on `vote` after holding the `partyid` variable constant. To do this, we need to include both `FT` and `partyid` in the same model; however, we cannot do this with the data in their current form, since the values of `partyid` do not vary across alternatives (which is required by the conditional logit specification).

The key to getting around these problems is to recognize that variables which do not vary across alternatives must be allowed to exert a separate impact on each alternative (as they do in the MNL model). Including separate constant terms – analogous to "fixed effects" – for the alternatives turns out to be the key to doing this. This can be done by simply creating indicator variables for each choice; we can then interact each of the observation-specific covariates with those alternative-specific constants, and include both the direct and the interactive terms in a conditional logit model:

```
. gen clintondummy=(candid==2)

. gen perotdummy=(candid==3)
```

```
. gen PIDxClinton = partyid * clintondummy

. gen PIDxPerot = partyid * perotdummy

. clogit vote clintondummy perotdummy PIDxClinton PIDxPerot, group(caseid)

Conditional (fixed-effects) logistic regression   Number of obs   =       4419
                                                  LR chi2(4)      =    1069.56
                                                  Prob > chi2     =     0.0000
Log likelihood = -1083.4749                       Pseudo R2       =     0.3305
------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
clintondummy |   4.842495   .2373171    20.41   0.000     4.377362    5.307628
  perotdummy |   1.815236    .245572     7.39   0.000     1.333923    2.296548
 PIDxClinton |  -1.163168   .0545618   -21.32   0.000    -1.270107   -1.056229
   PIDxPerot |  -.4826514   .0475563   -10.15   0.000      -.57586   -.3894427
------------------------------------------------------------------------------
```

Note that these results are exactly the same as those given by the MNL model, above; the coefficients, standard errors, and log-likelihood are identical to the MNL. The only difference is in the goodness-of-fit statistics.

Using alternative-specific dummies – and their interactions with observation-specific covariates – thus lets us estimate a MNL model with "expanded" (CL) data. Importantly, this also allows us to add variables that vary by choice (such as `FT`) to the model:

```
. clogit vote FT clintondummy perotdummy PIDxClinton PIDxPerot, group(caseid)

Conditional (fixed-effects) logistic regression   Number of obs   =       4419
                                                  LR chi2(5)      =    1764.49
                                                  Prob > chi2     =     0.0000
Log likelihood =  -736.0092                       Pseudo R2       =     0.5452
------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          FT |   .0629875   .0032175    19.58   0.000     .0566813    .0692937
clintondummy |   2.812737   .2687999    10.46   0.000     2.285899    3.339576
  perotdummy |   .9435437   .2856252     3.30   0.001     .3837286    1.503359
 PIDxClinton |  -.6318723    .062255   -10.15   0.000    -.7538899   -.5098548
   PIDxPerot |  -.1921175    .057032    -3.37   0.001    -.3038981   -.0803369
------------------------------------------------------------------------------
```

In this model, we can simultaneously estimate the influence of choice-specific and observation-specific covariates on the probability of a particular choice.

Conditional Logit: Another Example

Maltzmann and Wahlbeck (1996) study opinion assignment in the Rehnquist Court.[2] Their data are majority opinion assignments by Chief Justice Rehnquist ($N = 316$, with between 4 and 9 possible choice alternatives on each, for 2213 total lines of data). Their independent variables are *Ideology*, *Equity*, *Expertise*, and *Efficiency*, plus a *CJ Dummy* of the potential assignee (these are all choice-specific variables), plus *Consensus*, *MWC*, *Importance* and *End-of-Term* variables (which are case-specific).

Their analysis points out something else: that you don't necessarily need to interact the observation-specific variables with dummies for the choices...

- So long as they're interacted with *some* variable that varies across alternatives, the model will be identified.
- Here, the theory (such as it is) suggested that they should interact a number of their variables with *ideology*...

Next week, we'll discuss at length how to interpret and discuss MNL and CL models.

[2] M&W call it a "multinomial logit" model and a "discrete choice" model, but it's a conditional logit, with interaction terms to take into account case-specific variables...